JOURNAL OF COMBINATORIAL THEORY 10, 135–142 (1971)

A Context-Free Language and Enumeration Problems on Infinite Trees and Digraphs

W. KUICH

IBM Laboratory, Vienna, Austria

Communicated by J. Riordan

Received June 19, 1969

There are given: an infinite tree T_r, an infinite digraph D_r, and a context-free grammar G_r. Then the relations between four problems are investigated: in T_r, counting subtrees containing k edges (one of them fixed); in D_r, counting paths of length $2rk + 1$; in the language generated by G_r, counting the words of length $2rk + 1$; and in T_2, counting the subtrees satisfying a certain condition.

1. STATEMENT OF THE PROBLEM

PROBLEM A. We are given the infinite tree T_r with vertex set

$$V_r = \{v(1, 1)\} \cup \{v(i,j) \mid i \geqslant 2;\ 1 \leqslant j \leqslant r^{i-2}\}$$

and edge set

$$E_r = \{(v(1, 1), v(2, 1))\}$$
$$\cup \{(v(i,j+1), v(i+1, rj+k)) \mid i \geqslant 2; 0 \leqslant j \leqslant r^{i-2} - 1;\ 1 \leqslant k \leqslant r\}$$

The tree T_2 is given in Figure 1.

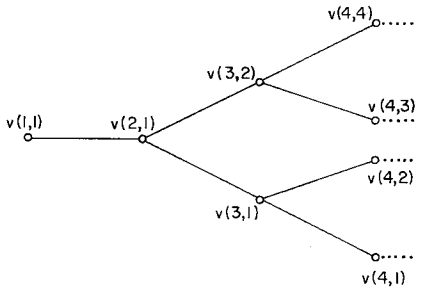

FIG. 1. The tree T_2.

We want to calculate the number $a_r(k)$ of subtrees of T_r that contain exactly k edges including the edge $(v(1, 1), v(2, 1))$.

PROBLEM B. We are given the acyclic infinite digraph D_r with point set

$$P_r = \{p(i,j) \mid i \geqslant 1; 1 \leqslant j \leqslant 2r^{i-1}\}$$

and line set

$$\begin{aligned} L_r = &\{\langle p(i, 2j+1), p(i+1, 2rj+1)\rangle \mid i \geqslant 1; 0 \leqslant j \leqslant r^{i-1} - 1\} \\ &\cup \{\langle p(i,j), p(i, j+1)\rangle \mid i \geqslant 1; 1 \leqslant j \leqslant 2r^{i-1} - 1\} \\ &\cup \{\langle p(i+1, 2rj), p(i, 2j)\rangle \mid i \geqslant 1; 1 \leqslant j \leqslant r^{i-1}\}. \end{aligned}$$

The digraph D_2 is given in Figure 2.

We want to calculate the number $b_r(k)$ of (directed) paths of length k from $p(1, 1)$ to $p(1, 2)$.

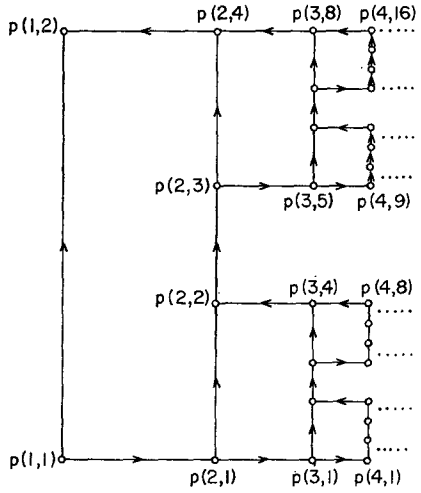

FIG. 2. The digraph D_2.

To increase readability we mention some facts of the theory of context-free languages.

An alphabet is a finite non-empty set. The set of all words, including the empty word ϵ, over an alphabet Σ is denoted by Σ^*.

A context-free grammar is a 4-tuple $G = (V, \Sigma, P, \sigma)$, where (i) V is an alphabet (the vocabulary); (ii) $\Sigma \subseteq V$ is an alphabet (the terminals); (iii) P is a finite set of productions of the form $\xi \to v$, where ξ is in $V - \Sigma$ and v is in V^*; (iv) σ is in $V - \Sigma$ (the start symbol).

Elements of $V - \Sigma$ are called non-terminals.
For x, y in V^*, write

$$x \underset{l}{\Rightarrow} y$$

if there exist ξ in $V - \Sigma$, u in Σ^*, v, w in V^* such that $x = u\xi v$, $y = uwv$ and $\xi \to w$ is in P. For x, y in V^*, write

$$x \overset{*}{\underset{l}{\Rightarrow}} y$$

if there exist $z_0, \ldots, z_r$ such that $z_0 = x$, $z_r = y$, and

$$z_i \underset{l}{\Rightarrow} z_{i+1}$$

for $i = 0, \ldots, r - 1$. Such a sequence $z_0, \ldots, z_r$ of words is called a leftmost derivation or a leftmost generation (from z_0 to z_r) of length r and is written

$$z_0 \underset{l}{\Rightarrow} \cdots \underset{l}{\Rightarrow} z_r .$$

A subset L of Σ^* is called a context-free language if $L = L(G)$ for some grammar $G = (V, \Sigma, P, \sigma)$, where

$$L(G) = \{w \text{ in } \Sigma^* \mid \sigma \overset{*}{\underset{l}{\Rightarrow}} w\}.$$

$L(G)$ is said to be the language generated by G.

A grammar $G = (V, \Sigma, P, \sigma)$ is said to be ambiguous if there is some word in $L(G)$ generated by at least two different leftmost derivations from σ. A grammar which is not ambiguous is said to be unambiguous.

Let L be a language and let $u(n)$ be the structure function of L, i.e., $u(n)$ is the number of distinct words of length n contained in L. Then the function $f(z)$ of the complex variable z

$$f(z) = \sum_{n=1}^{\infty} u(n)\, z^n$$

is called the structure generating function of L.

Assuming that the unambiguous context-free grammar $G = (V, \Sigma, P, \gamma_1)$ with non-terminals $V - \Sigma = \{\gamma_1, \ldots, \gamma_k\}$ and terminals $\Sigma = \{x_1, \ldots, x_l\}$ contains no productions of the form $\gamma_i \to \gamma_j$ or $\gamma_i \to \epsilon$, it is possible to construct a system of equations whose unique solution is the structure generating function of $L(G)$.

Let $\phi_{i,1},\ldots,\phi_{i,m_i}$ be all the elements of V^* such that $\gamma_i \to \phi_{i,j}$, $(1 \leqslant j \leqslant m_i)$, is an element of P and write

$$\gamma_i = \phi_{i,1} + \cdots + \phi_{i,m_i}$$

for all γ_i in $V - \Sigma$. Replacement of each occurrence of γ_j and x_m by the power series y_j and the complex variable z, respectively, yields a system of equations

$$y_i = H_i(y_1,\ldots,y_k\,;z) \qquad (1 \leqslant i \leqslant k).$$

which has a unique solution

$$y_i = f_i(z), \qquad f_i(0) = 0 \qquad (1 \leqslant i \leqslant k).$$

Besides $f_1(z)$ is the structure generating function of $L(G)$ (compare Theorem 2 of Kuich [3]).

PROBLEM C. We are given the unambiguous context-free grammar

$$G_r = (V, \Sigma, P_r, \sigma),$$

where $V = \{\sigma\} \cup \Sigma$, $\Sigma = \{a, b\}$, and $P_r = \{\sigma \to (a\sigma)^r a, \sigma \to b\}$ (Note that x^r, x a symbol, means the concatenation of r symbols x.)

We want to calculate the structure function $c_r(k)$ of $L(G_r)$.

2. CALCULATION OF THE GENERATING FUNCTIONS

By Kuich [3], the structure generating function

$$C_r(z) = \sum_{k=1}^{\infty} c_r(k)\, z^k$$

is the unique solution of the equation

$$y_r(z) = z^{r+1} y_r(z)^r + z, \tag{1}$$

with $y_r(0) = 0$.

The transformation

$$w(x, r) = y_r(z)/z, \qquad x = z^{2r} \qquad \text{with} \quad w(0, r) = 0,$$

yields the equation

$$1 - w(x, r) + x[w(x, r)]^r = 0,$$

which is solved by Pólya and Szegö [5, p. 125, No. 211]. Retransformation yields

$$c_r(k) = \begin{cases} \dfrac{(rn)!}{n!\,[(r-1)n+1]!} & (k = 2rn+1;\ n \geqslant 0), \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

In the case $r = 2$ it is possible to solve (1) explicitly:

$$C_2(z) = \frac{1 - \sqrt{1 - 4z^4}}{2z^3}. \tag{3}$$

Kuich [3] showed how to associate a digraph with a given grammar such that the number of paths of length k from the initial point to the final point is equal to the number of words of length k generated by the grammar.

Since D_r is associated with G_r in the manner described above,

$$B_r(z) = C_r(z), \tag{4}$$

where $B_r(z)$ is the generating function

$$B_r(z) = \sum_{k=1}^{\infty} b_r(k)\, z^k.$$

The next step is to relate $B_r(z)$ and $A_r(z)$, the generating function of the numbers $a_r(k)$.

Each path in D_r from $p(1, 1)$ to $p(1, 2)$ is uniquely identified by the lines of the form

$$\langle p(i, 2j+1), p(i+1, 2rj+1)\rangle, \qquad i \geqslant 1;\quad 0 \leqslant j \leqslant r^{i-1} - 1$$

that are contained in it.

By the one-to-one mapping of these lines onto the edges of the tree T_r:

$$\begin{aligned} &\langle p(i, 2rj + 2(k-1) + 1), p(i+1, 2r^2 j + 2r(k-1) + 1)\rangle \\ &\quad \leftrightarrow (v(i, j+1), v(i+1, rj+k)) \\ &\quad i \geqslant 2, \qquad 0 \leqslant j \leqslant r^{i-2} - 1, \qquad 1 \leqslant k \leqslant r \end{aligned}$$

and

$$\langle p(1, 1), p(2, 1)\rangle \leftrightarrow (v(1, 1), v(2, 1))$$

each path of length $k = 2rn + 1$ $(n \geqslant 1)$ from $p(1, 1)$ to $p(1, 2)$ in D_r is mapped on a subtree of T_r that contains exactly n edges including $(v(1, 1), v(2, 1))$.

Hence

$$a_r(n) = b_r(2rn + 1), \qquad n \geqslant 1$$

and

$$A_r(z) = \sum_{n=1}^{\infty} \frac{(rn)!}{n!\,[(r-1)n+1]!}\, z^n. \tag{5}$$

The result (5) is a generalization of a result achieved by Izbicki [2], who calculated $a_2(n)$ with an entirely different method.

In case $r = 2$,

$$A_2(z) = c(z) - 1, \tag{6}$$

where $c(z)$ is the generating function of Catalan numbers

$$c_n = \frac{1}{n+1}\binom{2n}{n}.$$

Professor Riordan mentioned to me that $a_r(n)$ is the same as what Riordan [6] and Carlitz [1] denoted by $g_{n-1}(r+1)$, which is the number of line chromatic planted trees with $r+1$ line colors, a given color on the stem and n lines. A tree is line chromatic when no two adjacent lines have the same color.

3. A Generalization of Problem A

Problem D. We assign natural numbers r and s to the edges of the tree T_2 in the following manner:

r is assigned to $(v(1, 1), v(2, 1))$ and to $(v(i, j+1), v(i+1, 2j+2))$, $i \geqslant 2;\ 0 \leqslant j \leqslant r^{i-2} - 1$.

s is assigned to $(v(i, j+1), v(i+1, 2j+1))$, $i \geqslant 2;\ 0 \leqslant j \leqslant r^{i-2} - 1$.

Hence the assignment of the numbers r and s is: r to the first edge. and after every bifurcation, r to the upper edge, s to the lower.

We want to calculate the number $d(k; r, s) = d(k)$ of subtrees of T_2 that contain the edge $(v(1, 1), v(2, 1))$ and satisfy the following condition: the sum of the numbers assigned to the edges of the subtree equals k.

To find

$$D(z; r, s) = \sum_{k=1}^{\infty} d(k)\, z^k$$

we proceed as before.

Let G be the context-free grammar

$$G = (V, \Sigma, P, \sigma_1),$$

where

$$V = \{\sigma_1, \sigma_2\} \cup \Sigma, \qquad \Sigma = \{a, b\}$$

and

$$P = \{\sigma_1 \to a\sigma_2 a^{4r-3}\sigma_1 a,\ \sigma_1 \to b,\ \sigma_2 \to a\sigma_2 a^{4s-3}\sigma_1 a,\ \sigma_2 \to b\}.$$

Let $f_1(k)$ be the structure function of G. Then the structure generating function

$$F_1(z) = \sum_{k=1}^{\infty} f_1(k)\, z^k$$

is equal to the unique solution $y_1(z)$, $y_1(0) = 0$ of the system of equations

$$y_1(z) = z^{4r-1} y_1(z)\, y_2(z) + z,$$
$$y_2(z) = z^{4s-1} y_1(z)\, y_2(z) + z.$$

Hence

$$F_1(z) = \frac{1 - z^{4r} + z^{4s} - \sqrt{(1 - z^{4r} + z^{4s})^2 - 4z^{4s}}}{2z^{4s-1}}.$$

As before, $f_1(k)$ equals the number of paths from the initial point $p(1, 1)$ to the final point $p(1, 2)$ of the digraph D, associated with the grammar G. D is drawn in Figure 3. Note that $f_1(k) = 0$ if $k \neq 4(k_1 r + k_2 s) + 1$, $k_1, k_2 \geqslant 0$.

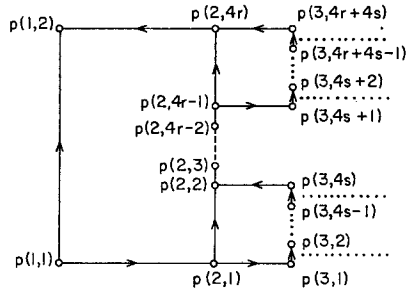

FIG. 3. The digraph D.

Each path from $p(1, 1)$ to $p(1, 2)$ in D of length $4(k_1 r + k_2 s) + 1$, $k_1 \geqslant 1$, $k_2 \geqslant 0$ is mapped one-to-one on a subtree of the tree T_2 that contains the edge $(v(1, 1), v(2, 1))$ and has the property that the sum of the numbers, assigned to its edges, equals $k_1 r + k_2 s$.

Hence

$$d(k_1r + k_2s) = f_1(4(k_1r + k_2s) + 1) \qquad k_1 \geqslant 1, \quad k_2 \geqslant 0.$$

Since $f_1(1) = 1$,

$$D(z; r, s) = \frac{F_1(\sqrt[4]{z})}{\sqrt[4]{z}} - 1$$

or

$$D(z; r, s) = \frac{1 - z^r + z^s - \sqrt{(1 - z^r + z^s)^2 - 4z^s}}{2z^s} - 1.$$

ACKNOWLEDGMENT

Thanks are due to Professor John Riordan for several remarks.

REFERENCES

1. L. CARLITZ, A note on the enumeration of line chromatic trees, *J. Combinatorial Theory* **6** (1969), 99–101.
2. H. IZBICKI, Über Unterbäume eines Baumes, *Monatsh. Math.* **74** (1970), 56–62.
3. W. KUICH, On the entropy of context-free languages, *Information and Control* **16** (1970), 173–200.
4. W. KUICH, Languages and the enumeration of planted plane trees, *Indag. Math.* **32** (1970), 268–280.
5. G. PÓLYA AND G. SZEGÖ, "Aufgaben und Lehrsätze aus der Analysis," Vol. I. Springer, Berlin, 1925; reprinted Dover, New York, 1945.
6. J. RIORDAN, The number of labeled, colored and chromatic trees, *Acta Math.* **97** (1957), 211–225.